Notes on Bates et al. (2021)

Ryan Giordano

June 9, 2023

1 Setup and problem statement

This paper takes as given a black-box algorithm $f(\cdot)$ that operates on pairs $Z_n := (X_n, Y_n)$, producing a prediction $f(X^*) := Y^*$, which we hope satisfies $\hat{Y}^* \approx Y^*$ (though we assume nothing of the form). The algorithm $f(\cdot)$ is typically constructed from a training set, which we will ignore.

We have some calibration set, $\mathcal{Z} := \{\tilde{Z}_1, \ldots, \tilde{Z}_N\}$, which we would like to use to form *interval-valued* predictions for Y^* from Y^*. That is, we will use the observations in $\mathcal{Z}$ to form a (random) set-valued function, $\mathcal{C}(X^*)$. (We could write $\mathcal{C}(X^*|\mathcal{Z})$, to emphasize the dependence on the calibration set but that would get tedious.)

I'll use $\mathcal{S}(\cdot)$ for the mapping and $\mathcal{S}$ for sets.

How to choose the mapping $\mathcal{C}(\cdot)$ to have desirable properties? We define a family of candidate sets, and a loss function describing what a "good" set looks like. Specifically, let's take

- A nested one-dimensional family of set functions, $\mathcal{C}_\lambda(\cdot)$ such that bigger λ results in bigger sets:

$$\lambda_1 < \lambda_2 \Rightarrow \mathcal{C}_{\lambda_1}(X) \subset \mathcal{C}_{\lambda_2}(X) \text{ for all } X.$$

- A loss function $\mathcal{L}(Y, \mathcal{S})$ which increases as sets get smaller:

$$\mathcal{S} \subset \mathcal{S} \Rightarrow \mathcal{L}(Y, \mathcal{S}) \geq \mathcal{L}(Y, \mathcal{S}').$$

Tension: We want small sets (small λ), but also want small $\mathcal{L}$ (big λ). This paper is all about how to choose λ to balance these desiderata with statistical guarantees.

How to choose sets? The paper points out that for models which return an estimate of the relative probabilities of Y, one can correspondingly impute the probability of $\mathcal{L}(Y, \mathcal{S}+\{Y\})$, and doing so greedily design a nested family of sets which is optimal if your model is correct.

Problem: How to use the calibration set $\mathcal{Z}$ to choose $\hat{\lambda}$ so that the loss $\mathcal{L}(Y^*, \mathcal{C}_{\hat{\lambda}}(X^*)$ is "probably" small, where "probably" accounts for randomness in both $\mathcal{Z}$ and in (X^*, Y^*)?

2 Running example

The paper really shines in its creation of meaningful loss functions for complex settings. But I think this short exposition will be better served by a much more familiar example which connects more closely to more standard conformal inference.

Suppose $Y^* \in \mathbb{R}$ (so it's like a regression problem), and we want to flag Y^* that we think are too big. So what we want is a one-sided confidence interval for Y^*:

$$\begin{aligned}\mathcal{C}_\lambda(X^*) &= \{Y : Y \leq \hat{Y} + \lambda\} \\ \mathcal{L}(Y, \mathcal{S}) &= \mathbb{I}\left(\lambda < Y\right).\end{aligned}$$

So we incur a loss of one if we fail to cover. This satisfies the above definitions. □

Note that classical conformal inference gives a solution to this problem using the non-conformity score $Y - \hat{Y}$, setting λ to be an appropriate quantile, controlling

$$\underset{\mathcal{Z}, Z^*}{\mathbb{P}}\left(\mathcal{L}(Y^*, \mathcal{C}_{\hat{\lambda}}(X^*) \geq \alpha\right)$$

This paper will do something a little different, though we'll try to connect the two at the end.

3 This paper's solution

3.1 Define probably small

First, the paper defines "probably small" as

$$\underset{\mathcal{Z}}{\mathbb{P}}\left(\underset{Z^*}{\mathbb{E}}\left[\mathcal{L}(Y^*, \mathcal{C}_{\hat{\lambda}}(X^*)\right] \leq \alpha\right) =: \underset{\mathcal{Z}}{\mathbb{P}}\left(\mathscr{R}(\hat{\lambda}) \leq \alpha\right) \geq 1 - \delta$$

The expectation will be called the "risk," and we'll use it enough to give it a name:

$$\mathscr{R}(\mathcal{C}_\lambda) = \mathscr{R}(\lambda) = \mathop{\mathbb{E}}_{Z^*}\left[\mathcal{L}(Y^*, \mathcal{C}_\lambda(X^*)\right].$$

Note that this is different than standard conformal prediction (see above), though the two are related. Vovk (2012) calls this "training conditional" validity because the inner expectation has the calibration set fixed. It is also a guarantee similar to "tolerance regions" (Krishnamoorthy and Mathew, 2009), which we will discuss if we have time.

Example. In our running example,

$$\mathscr{R}(\lambda) = \mathop{\mathbb{E}}_{Z^*}\left[\mathbb{I}\left(Y^* - \hat{Y} > \lambda\right)\right] = \mathop{\mathbb{P}}_{Z^*}\left(Y^* - \hat{Y} > \lambda\right).$$

That is, $\mathscr{R}(\lambda)$ is just one minus the distribution function of the non-conformity score, evaluated at λ. So we want to choose $\hat{\lambda}$ so that it is larger than the true $1-\alpha$ quantile, with probability (in the calibration set) at least $1-\delta$. This amounts to constructing a good estimate of the distribution function — which we can do using the empirical distribution function on the calibration set. □

3.2 Control the risk using the calibration set

How to control $\mathscr{R}(\cdot)$ using $\mathcal{Z}$? If you knew the risk function, you would simply take

$$\lambda^* := \inf\{\lambda : \mathscr{R}(\lambda) \leq \alpha\} \text{ and } \delta = 0.$$

But we don't, so we have to estimate $\mathscr{R}(\cdot)$ using $\mathcal{Z}$. Note that we're going to both estimate the function $\lambda \mapsto \mathscr{R}(\lambda)$, and then search over our estimate to pick a λ. You might think that would require a uniform bound on the accuracy of our approximation, but it won't, due to a clever expolitation of monotonicity of the loss.

The authors assume that you can form a one-sided lower confidence region for $\mathscr{R}(\lambda)$, for any λ (pointwise). That is, that you can find an upper confidence bound (UCB) $\hat{\mathscr{R}}^+(\lambda)$ such that

$$\mathop{\mathbb{P}}_{\mathcal{Z}}\left(\mathscr{R}(\lambda) \leq \hat{\mathscr{R}}^+(\lambda)\right) \geq 1-\delta.$$

This UCB is all you need! There are lots of ways to construct it, using concentration inequalities, or even asymptotics (we will talk later). But once you have it, you can take

$$\hat{\lambda} := \inf\{\lambda : \hat{\mathscr{R}}^+(\lambda') < \alpha \text{ for all } \lambda' > \lambda\}.$$

Here's their proof that this works (in the case that $\mathscr{R}(\lambda)$ is continuous):

- Suppose we picked $\hat{\lambda}$ "too small:" $\hat{\lambda} < \lambda^*$, and $\mathscr{R}(\hat{\lambda}) > \mathscr{R}(\lambda^*) = \alpha$. We failed to achieve our bound — the risk at $\hat{\lambda}$ is too high.
- But we chose $\hat{\lambda}$ so that $\hat{\lambda} < \lambda^* \Rightarrow \hat{\mathscr{R}}^+(\lambda^*) < \alpha = \mathscr{R}(\lambda^*)$. In other words, the risk $\mathscr{R}(\lambda^*)$ is outside its confidence interval $(-\infty, \hat{\mathscr{R}}^+(\lambda^*))$.
- By construction $\hat{\mathscr{R}}^+(\cdot)$, this can happen with probability at most $1-\delta$. Therefore we fail to control risk with probability no more than $1-\delta$.

3.3 Choose an upper confidence bound

We now need only choose an UCB. Note that we can compute

$$\hat{\mathscr{R}}(\lambda) := \frac{1}{N}\sum_{n=1}^{N} \mathcal{L}(Y_n, \mathcal{C}_\lambda(X_n)).$$

For any λ, $\hat{\mathscr{R}}(\lambda)$ is an unbiased estimate of $\mathscr{R}(\lambda)$. Different concentration results of $\hat{\mathscr{R}}(\lambda)$ to its mean $\mathscr{R}(\lambda)$ can give a family of UCB.

The simplest example is Hoeffding in the case that $\mathcal{L}(\cdot) \in [0,1]$:

$$\begin{aligned}
\mathop{\mathbb{P}}_{\mathcal{Z}}\left(\hat{\mathscr{R}}(\lambda) - \mathscr{R}(\lambda) < -t\right) \leq \exp\left(-2Nt^2\right) = \delta \Leftrightarrow \\
\mathop{\mathbb{P}}_{\mathcal{Z}}\left(\mathscr{R}(\lambda) > \hat{\mathscr{R}}(\lambda) + t\right) \leq \exp\left(-2Nt^2\right) = \delta \Leftrightarrow \\
\hat{\mathscr{R}}^+(\lambda) := \hat{\mathscr{R}}(\lambda) + \sqrt{\ln\delta/(-2N)}.
\end{aligned}$$

This is loose, but easy to understand. For bounded losses they actually recommend the Waudby-Smith-Ramdas bound, which is based on a maximal inequality for martingales and is better able to take into account different variances for different λ.

For unbounded losses, you need to assume something. They consider a Pinelis-Utev inequality, but also consider asymptotic normal bounds.

Again, the fact that $\mathscr{R}(\lambda)$ is monotonic eliminates the need to use bounds of the form $\sup_\lambda(\hat{\mathscr{R}}(\lambda) - \mathscr{R}(\lambda))$. Simple pointwise bounds are enough.

Example. In our example, we need to control

$$\frac{1}{N}\sum_{n=1}^{N}\mathbb{I}\left(Y_n - \hat{Y}_n > \lambda\right) - \mathscr{R}(\lambda),$$

where we used the fact that the risk is the expected loss. Note that this quantity is $1/N$ times a binomial random variable with probability $\mathscr{R}(\lambda)$.

If we knew $\mathscr{R}(\lambda) = \rho$, then we can use the binomial distribution to find $t(\mathscr{R}(\lambda))$ such that

$$\mathop{\mathbb{P}}_{\mathcal{Z}}\left(\frac{1}{N}\sum_{n=1}^{N}\mathbb{I}\left(Y_n - \hat{Y}_n > \lambda\right) - \rho \geq t(\rho)\right) \geq 1 - \delta.$$

That is, we can form an exact, finite-sample test of the hypothesis $\mathscr{R}(\lambda) = \rho$. We can then invert the test to find a UCB:

$$\hat{\mathscr{R}}^{+}(\lambda) := \sup\{\rho : \text{We do not reject } \rho \text{ for the observed } \mathcal{Z}\}.$$

□

The paper generalizes the preceding idea with a number of different concentration inequalities.

4 Experiments

Their experiments exhibit a lot of clever loss functions, and show that their procedure results in reasonable (not too wide) intervals whose risk centers, roughly, on the target. Going through the experiments is probably too much for a short oral talk; check out the paper.

5 A classical version

Let's assume that each loss, $\lambda \mapsto \mathcal{L}(Y_n, \mathcal{C}_\lambda(X_n))$ is also monotonic (non-increasing).

An alternative to the present work would be to use the smallest λ needed to achieve a given loss on a *particular example* as a non-conformity score:

$$\ell_n := \ell(Z_n) := \inf\{\lambda : \mathcal{L}(Y_n, \mathcal{C}_\lambda(X_n)) \leq \alpha\}.$$

By monotonicity, if $\lambda \geq \ell(Z_n)$, then $\mathcal{L}(Y_n, \mathcal{C}_\lambda(X_n)) \leq \alpha$. So we want to choose $\hat{\lambda}$ to guarantee that $\ell(Z^*) \geq \hat{\lambda}$ with high probability. To acheive

this, we can take as $\hat{\lambda}$ the appropriate quantile of the non-conformity scores: $\hat{\lambda} := \ell_{(\lfloor \delta(N+1) \rfloor)}$. This would guarantee

$$\underset{\mathcal{Z},Z^*}{\mathbb{P}} \left(\mathcal{L}(Y^*, \mathcal{C}_{\hat{\lambda}}(Y^*) \leq \alpha \right) \geq \underset{\mathcal{Z},Z^*}{\mathbb{P}} \left(\ell(Z^*) \geq \hat{\lambda} \right) \geq 1 - \delta.$$

This is more in the spirit of Gupta, Kuchibhotla, and Ramdas (2022) (which the authors cite as the source of the idea of nested sets).

Example. In our example, the loss function is binary, so we can only control the probability that it's non-zero — we cannot target a loss equal to any $\alpha \in (0, 1)$. But the formalism works; for any $\alpha > 0$,

$$\ell(Z_n) = \inf\{\lambda : \mathcal{L}(Y_n, \mathcal{C}_\lambda(X_n)) \leq \alpha\} = \inf\{\lambda : \mathbb{I}\left(Y_n - \hat{Y} \geq \lambda\right) \leq \alpha\} = Y_n - \hat{Y}.$$

So $\hat{\lambda}$ is the $\lfloor \delta(N+1) \rfloor$-th quantile of the non-conformity scores, and this procedure returns the standard conformal one-sided interval. □

The nature of the bound is different, though as pointed out by Vovk (2012), one implies the other for bounded losses, and with different constants. But the computational requirements are different as well.

- The present paper requires computation of $\hat{\mathscr{R}}^+(\lambda)$ and then (roughly speaking) inversion of the map $\lambda \mapsto \hat{\mathscr{R}}^+(\lambda)$.
- The above approach requires inversion of the map $\lambda \mapsto \mathcal{L}(Y_n, \mathcal{C}_\lambda(X_n))$, but then only a quantile computation to get $\hat{\lambda}$.

I expect this proposal to require more computational effort, especially for complicated loss functions.

References

Bates, S. et al. (2021). "Distribution-free, risk-controlling prediction sets". In: *Journal of the ACM (JACM)* 68.6, pp. 1–34.

Gupta, C., A. Kuchibhotla, and A. Ramdas (2022). "Nested conformal prediction and quantile out-of-bag ensemble methods". In: *Pattern Recognition* 127, p. 108496.

Krishnamoorthy, K. and T. Mathew (2009). *Statistical tolerance regions: Theory, applications, and computation*. John Wiley & Sons.

Vovk, V. (2012). "Conditional validity of inductive conformal predictors". In: *Asian conference on machine learning*. PMLR, pp. 475–490.